Arquitetura de Gerenciamento Baseado em Intenções para Open RAN com Controle Dinâmico de Largura de Banda

João Vitor A. Garcês[1], João André C. Watanabe[1], Nicollas R. de Oliveira[1], Rodrigo S. Couto[2], Igor M. Moraes[1], Dianne S. V. de Medeiros[1], Diogo M. F. Mattos[1]

[1] LabGen/MídiaCom – TET/IC/PPGEET
Universidade Federal Fluminense - UFF
Niterói, RJ – Brasil

[2]Grupo de Teleinformática e Automação (GTA) - PEE/COPPE
Universidade Federal do Rio de Janeiro - UFRJ - Brazil

Abstract. *Quality of Service (QoS) is mandatory in next-generation mobile networks (beyond 5G and 6G) to support enhanced Mobile Broadband (eMBB), Ultra-Reliable Low-Latency Communication (URLLC), and Massive Machine-Type Communication (mMTC) in an Open Radio Access Network (Open RAN). However, the complexity of managing these networks increases the likelihood of configuration errors, leading to QoS degradation. This paper proposes an intent-based system where operators control resource allocation through natural language expressions translated into network policies for dynamic bandwidth control. The system is validated in an emulated environment, OpenFlow meters, and traffic queues. The results demonstrate efficient resource allocation and compliance with Service Level Agreements (SLAs), ensuring QoS and traffic prioritization for flows.*

Resumo. *A Qualidade de Serviço (QoS) é mandatória em redes móveis de próxima geração (além do 5G e 6G) para suportar serviços de banda larga móvel aprimorada (eMBB), comunicação ultraconfiável de baixa latência (URLLC) e comunicação massiva entre máquinas (mMTC) em uma Rede de Acesso via Rádio Aberta (Open RAN). Contudo, a complexidade na gestão dessas redes aumenta a probabilidade de erros de configuração que implicam degradação da QoS. Este artigo propõe um sistema baseado em intenções, no qual os operadores controlam a alocação de recursos por meio expressões em linguagem natural, traduzidas em políticas de rede para controle dinâmico de largura de banda. O sistema é validado em um ambiente emulado, medidores OpenFlow e filas de tráfego. Os resultados mostram a alocação eficiente de recursos e conformidade com os Acordos de Nível de Serviço (SLAs), garantindo a QoS e a priorização do tráfego para fluxos.*

1. Introdução

A crescente demanda por serviços móveis avançados, como veículos autônomos, realidade aumentada e Internet das Coisas (*Internet of Things* - IoT), impõe desafios significativos sobre as redes de próxima geração, além do 5G e 6G (*Beyond 5G and 6G* -

This work was funded by CNPq, CAPES, FAPERJ, RNP (Programa OpenRAN@Brasil, Processo MCTI Nº - A01245.014203/2021-14, e Programa de bolsa de Incentivo à Pesquisa), Niterói City Hall/FEC/UFF (Edital PDPA 2020) and INCT ICONIOT.

B5G/6G), para atender a requisitos variados de Qualidade de Serviço (*Quality of Service* - QoS) [Rezende et al., 2023]. Nas Redes de Acesso via Rádio Abertas (*Open Radio Access Networks* - Open RAN), que dissociam *hardware* e *software* por meio de interfaces abertas e interoperáveis, a gestão da QoS torna-se mais complexa devido à necessidade de coordenar componentes distribuídos da rede, como *fronthaul*, *midhaul* e *backhaul* [Filali et al., 2022]. A alocação eficiente de recursos é essencial para manter os Acordos de Nível de Serviço (*Service Level Agreements* - SLAs), considerando as categorias de serviço contratadas, como banda larga móvel aprimorada (*enhanced Mobile Broadband* - eMBB), comunicação ultra-confiável de baixa latência (*Ultra-Reliable Low-Latency Communication* - URLLC) e comunicação massiva entre máquinas (*massive Machine-Type Communication* - mMTC). A flexibilidade da Open RAN exige uma orquestração avançada para equilibrar esses aspectos em diversos serviços.

O fatiamento de rede (*network slicing*) é uma das tecnologias habilitadoras da Open RAN. Contudo, em um ambiente multi-fabricante, o fatiamento de rede enfrenta desafios adicionais para garantir isolamento e QoS diferenciada entre vários serviços. A gestão dinâmica da alocação de recursos, como os Blocos de Recursos (RBs) e a capacidade da nuvem de borda, torna-se cada vez mais complexa. A natureza distribuída da Open RAN exige coordenação precisa entre funções virtualizadas da rede para cumprir garantias de SLA ponta a ponta [Yeh et al., 2023]. A complexidade inerente à conformidade entre diferentes camadas da rede aumenta o risco de degradação da QoS entre fatias, tornando o desafio de manter a integridade do serviço ainda mais crítica.

O fatiamento de rede requer gestão cuidadosa para alcançar isolamento de recursos, elasticidade e personalização. Os operadores devem garantir que os recursos das fatias sejam adequadamente isolados, mantendo flexibilidade para se adaptar às demandas de serviço em tempo real [Abbas et al., 2021]. Redes Baseadas em Intenções (*Intent-Based Networking* - IBN) oferecem uma solução potencial, automatizando a configuração de fatias de rede com base em requisitos de QoS de alto nível [de Oliveira et al., 2024]. Ao interpretar as intenções definidas por operadores, um sistema baseado em IBN pode alocar recursos dinamicamente e otimizar o desempenho da rede, assegurando que diversos serviços coexistam na mesma infraestrutura física enquanto atendem seus SLAs específicos [Da Costa e Murillo, 2023].

Este artigo estende o sistema AGIR [Garcês et al., 2024, de Oliveira et al., 2024, de Oliveira et al., 2023] e propõe um sistema de gestão de fatias de rede baseado em intenções para ambientes Open RAN. Operadores de rede controlam a alocação de recursos por meio de intenções expressas em linguagem natural, que são convertidas em ações e aplicadas diretamente aos dispositivos de encaminhamento. A principal contribuição do artigo em relação a trabalhos anteriores é a utilização de uma interface baseada em intenções, facilitando a interação entre operadores e a rede, permitindo controle dinâmico de largura de banda para atender às demandas específicas de cada fatia e garantindo a conformidade com os SLAs. O artigo foca a garantia de QoS no núcleo da rede utilizando técnicas de virtualização e *softwarização*, permitindo a gestão eficiente de recursos e a adaptação automática às necessidades de cada serviço. Duas abordagens são utilizadas para o controle de largura de banda. Na primeira, são implementados medidores OpenFlow, enquanto na segunda são implementadas filas que obedecem à disciplina *Hierarchical Token Bucket* (HTB). A proposta destaca-se pela eficiência na alocação de re-

cursos devido à coordenação entre os planos de Gerência, Controle e Encaminhamento. A implementação da proposta em um ambiente emulado demonstra sua viabilidade prática em um cenário Open RAN.

No cenário de avaliação, as instâncias de Equipamentos de Usuário (UE) são emuladas utilizando a plataforma `srsRAN`. Os resultados mostram que, ao usar ambas as abordagens para o controle de banda em um cenário de alocação estático, assegura-se efetivamente a conformidade com os SLAs nas diferentes categorias de serviço do 5G (eMBB, URLLC e mMTC). Os medidores OpenFlow permitem limitação eficiente da largura de banda, enquanto as filas HTB garantem priorização de tráfego, essencial para manter a qualidade de serviço. Já em um cenário de alocação de banda dinâmica, as filas HTB superam superam os medidores OpenFlow. Enquanto os medidores OpenFlow promovem a justiça entre os fluxos, as filas HTB mantêm a priorização dos fluxos. Os experimentos emulados validam a capacidade do sistema de ajustar dinamicamente os recursos da rede, mesmo em cenários de fluxos simultâneos de diferentes serviços, garantindo que serviços de maior prioridade, como eMBB, recebam os recursos necessários sem comprometer os demais.

O restante do artigo está organizado da seguinte forma. A Seção 2 descreve os desafios do fatiamento de rede na RAN. A Seção 3 discute trabalhos relacionados. A Seção 4 detalha o sistema proposto, sua arquitetura e módulos. A Seção 5 descreve os cenários de avaliação em um ambiente emulado e discute os resultados experimentais. Por fim, a Seção 6 conclui o artigo e apresenta direções para pesquisas futuras.

2. Desafios no Domínio de Fatiamento de Rede da RAN

A implantação de fatiamento de rede na RAN apresenta diversos desafios que devem ser superados para garantir a eficácia dessa tecnologia, particularmente no contexto das redes de próxima geração (B5G/6G). Um dos principais desafios é a orquestração de fatias ponta a ponta. A orquestração exige que o fatiamento seja coordenado em toda a rede, desde a RAN até o núcleo e a rede de transporte. A orquestração envolve gerenciar recursos físicos e funções virtualizadas de rede de forma flexível e programável, garantindo que cada fatia tenha a quantidade correta de recursos para atender aos requisitos de QoS e SLA. Essa orquestração é crítica para o correto funcionamento da rede e é complexa devido à necessidade de integração entre diferentes camadas de rede e soluções de fornecedores [Afolabi et al., 2018].

A alocação eficiente de recursos na RAN é um obstáculo significativo na implantação do fatiamento de rede. Diferentemente do núcleo da rede, no qual a virtualização de funções apresenta um nível de maturidade mais elevado, a RAN possui limitações físicas, como a disponibilidade limitada de espectro e a necessidade de assegurar a coexistência de diferentes tecnologias de acesso via rádio. A limitação de espectro é exacerbada ao alocar canais de frequência dedicados a diferentes fatias da rede, uma vez que o particionamento rígido de recursos pode reduzir a eficiência da multiplexação [Rost et al., 2017]. Além disso, a coexistência de diferentes tecnologias de acesso na mesma infraestrutura física exige um planejamento cuidadoso e pode aumentar a complexidade na gestão das fatias.

A viabilidade econômica do fatiamento de rede é outro desafio crítico, especialmente para operadores de rede. Além das complexidades técnicas, é fundamental que o

fatiamento seja economicamente viável para justificar o investimento em novas arquiteturas de rede e virtualização. O fatiamento de rede pode gerar novas oportunidades de receita para as operadores ao permitir a criação de serviços personalizados para diferentes tipos de clientes sobre a infraestrutura já existente [Rost et al., 2017]. No entanto, a gestão eficiente de recursos entre diferentes fatias, enquanto se assegura flexibilidade e se maximiza a receita, requer o desenvolvimento de novos algoritmos de alocação de recursos e gestão de fatias.

O fatiamento na RAN também enfrenta desafios específicos em sua implementação para verticais. A alocação de recursos na RAN deve garantir isolamento adequado entre fatias, enquanto utiliza de forma eficiente o espectro compartilhado [Elayoubi et al., 2019]. Um dos principais problemas é a alocação eficiente de espectro, que evita interferências entre fatias e garante a conformidade com os SLAs de cada serviço. Além disso, a abertura da rede a terceiros, como operadores virtuais ou indústrias verticais, levanta questões sobre até que ponto esses atores podem monitorar e configurar suas fatias de rede.

A segurança e o isolamento das fatias são primordiais em ambientes multifabricantes, particularmente em redes 5G que suportam verticais. É crucial garantir que cada fatia esteja devidamente isolada para evitar interferências ou vazamentos de dados entre diferentes serviços. O isolamento das fatias em diferentes camadas da rede (RAN, núcleo e transporte) é necessário para proteger serviços críticos, como IoT para Redes Elétricas Inteligentes (*smart grids*), de congestionamentos ou falhas de segurança em outras fatias da rede [Faruque, 2021].

Os desafios enfrentados na implantação de fatiamento de rede em redes móveis, particularmente no contexto de redes B5G/6G, estão relacionados à complexidade de configurar múltiplas fatias de rede, cada uma com diferentes requisitos de QoS e SLA, em um ambiente de recursos compartilhados e limitados. A necessidade de coordenar componentes distribuídos, como a RAN, o núcleo da rede e as funções virtualizadas da rede, amplifica o risco de erros de configuração e conflitos de políticas quando gerenciados manualmente. Assim, a adoção de mecanismos automatizados de configuração, como soluções inteligentes de orquestração e IBN, torna-se essencial para garantir consistência na configuração, mitigar conflitos e assegurar que as fatias de rede operem de forma eficiente, respeitando os acordos de nível de serviço estabelecidos. Os sistemas baseados em intenções podem integrar o processamento de linguagem natural para permitir aos operadores o monitoramento e a configuração da rede de forma intuitiva. Dessa forma, reduz-se a necessidade de conhecimentos técnicos avançados, uma vez que o sistema tem capacidade de traduzir intenções de alto nível em configurações técnicas complexas, alinhadas às políticas de rede e aos requisitos de segurança previamente estabelecidos.

3. Trabalhos Relacionados

A orquestração do fatiamento de rede baseada em intenções tem emergido como uma abordagem promissora para lidar com a crescente complexidade das redes B5G/6G, permitindo o gerenciamento dinâmico de recursos e a garantia da QoS em diferentes cenários. Zhang *et al.* apresentam uma abordagem de orquestração de fatias na RAN baseada em intenções utilizando um algoritmo de aprendizado por reforço profundo multiagente (*Multi-Agent Deep Q-Network* - MA-DQN) [Zhang et al., 2023]. A proposta intro-

duz o conceito de Grau de Satisfação de SLA (*SLA Satisfaction Degree* - SSD) para medir a satisfação dos usuários e utiliza uma função de recompensa orientada por intenções para alocar dinamicamente blocos de recursos (RBs). Cada agente (usuário) otimiza a alocação de recursos com base em intenções específicas de QoS, melhorando a eficiência no uso dos recursos e o desempenho do sistema em termos de taxa de dados e cumprimento de SLAs. Os resultados das simulações mostram que o sistema supera os algoritmos tradicionais baseados em DQN, garantindo melhor convergência e alocação de recursos mais eficiente em cenários com múltiplos usuários.

Outros estudos investigam abordagens baseadas em aprendizado de máquina para otimização e controle de RAN. Habib *et al.* propõem um método de controle e orquestração inteligente orientado por intenções, utilizando aprendizado por reforço hierárquico para gerenciar rApps e xApps [Habib et al., 2023]. O método utiliza uma arquitetura de dois níveis composta por um meta-controlador e um controlador, com o objetivo de orquestrar múltiplas rApps ou xApps de acordo com as intenções do operador, otimizando indicadores como vazão, eficiência energética e latência.

GANSO é um arcabouço projetado para automatizar a criação e a configuração de fatias de rede em infraestruturas de rede definida por *software* (*Software-Defined Networking* - SDN), com foco em redes de transporte que conectam nuvens de borda e nuvens centrais [Infiesta et al., 2020]. O arcabouço utiliza modelos genéricos de fatias (*Generic Slice Templates* - GSTs) para mapear os parâmetros necessários para a criação das fatias. GANSO é implementado como uma aplicação do controlador de rede ONOS (*Open Networking Operating System*) e se comunica via API REST para configurar fatias de alta disponibilidade e desempenho. Abbas *et al.* desenvolveram um sistema de fatiamento de rede baseado em intenções, capaz de gerenciar eficientemente os recursos do núcleo e da RAN [Abbas et al., 2020]. O sistema automatiza a criação, configuração e monitoramento de fatias, utilizando IBN para traduzir intenções em políticas e orquestrá-las com o OSM (*Open Source MANO*) e o controlador de rede FlexRAN. Uma rede generativa adversária (*Generative Adversarial Network* - GAN) é empregada para prever estatísticas de uso de recursos, auxiliando na aceitação de fatias, escalonamento de recursos, recuperação de falhas e gerenciamento de recursos.

Aklamanu *et al.* apresentam um arcabouço baseado em intenções para gerenciar fatias de rede 5G, simplificando e automatizando a composição e o provisionamento de serviços [Aklamanu et al., 2018]. O sistema reduz a complexidade para operadores e inquilinos, permitindo a implantação de fatias em menos de dois minutos. Além disso, o arcabouço ajusta automaticamente as configurações conforme os SLAs, eliminando a necessidade de reconfigurações manuais para cada solicitação. Giorgetti *et al.* propõem uma solução baseada em SDN para garantir isolamento de desempenho entre fatias de rede, com foco em largura de banda e latência [Giorgetti et al., 2021]. Utilizando o protocolo OpenFlow, a solução é implementada no controlador ONOS e emprega medidores e filas de transmissão para controlar o tráfego de cada fatia. Além de assegurar o isolamento de desempenho, o sistema permite o empréstimo de largura de banda entre fatias quando há capacidade ociosa, otimizando o uso dos recursos.

Os trabalhos relacionados apresentam diversas abordagens para orquestração de redes e gerenciamento de recursos baseados em intenções. Em contraste, a solução proposta neste artigo ajusta dinamicamente a largura de banda e prioriza serviços com base

em intenções expressas em linguagem natural. Essa abordagem oferece maior flexibilidade e facilidade de uso na gestão de fatias em redes B5G/6G, aprimorando a eficiência operacional e a adaptabilidade do sistema.

4. Sistema Proposto de Gerenciamento de Rede Baseado em Intenções

A proposta estende o sistema AGIR (Agilidade no Gerenciamento baseado em Intenções para Refinamento de Níveis de Serviço) [de Oliveira et al., 2024, de Oliveira et al., 2023]. A proposta automatiza a configuração de políticas na RAN através de um assistente conversacional que interpreta intenções expressas em linguagem natural, garantindo o cumprimento dos parâmetros definidos nos SLAs.

A Figura 1 mostra a arquitetura multiplano na qual o sistema proposto está inserido. A arquitetura é composta pelos Planos de Gerenciamento, Controle e Encaminhamento. O Plano de Gerenciamento processa as intenções do operador e as transmite para

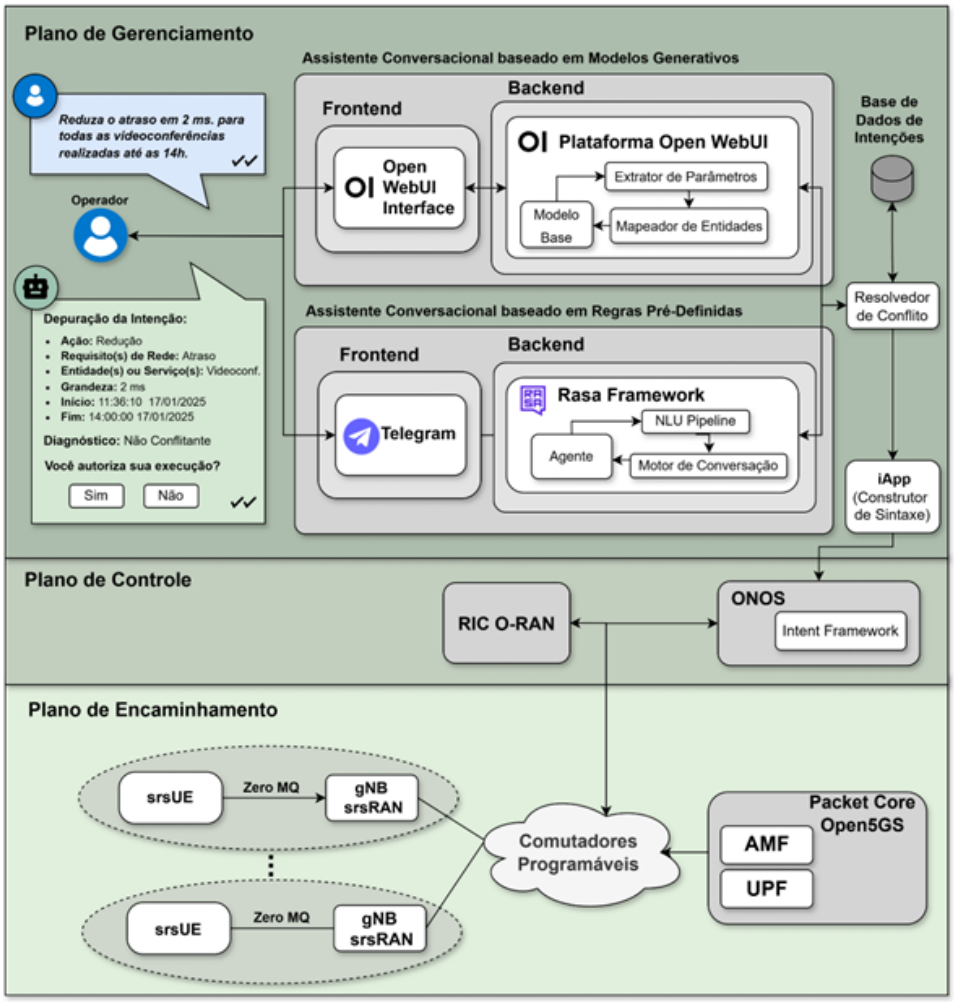


Figura 1. Arquitetura modular do sistema proposto captura e processa intenções expressas em linguagem natural por meio de interações com operadores via assistentes conversacionais. As intenções capturadas são traduzidas para uma sintaxe compreensível pelo controlador e validadas em relação aos possíveis conflitos com o estado atual da rede. Caso aprovada, a intenção é mapeada em instruções aplicáveis e executada na rede.

o Plano de Controle, onde são traduzidas e aplicadas no Plano de Encaminhamento. Assim, o Plano de Controle instrui o Plano de Encaminhamento a ajustar o fluxo de dados e a alocação de recursos, garantindo a implantação precisa das decisões de controle.

O **Plano de Gerenciamento** é responsável pela interface de gerenciamento e por traduzir as intenções do operador em ações executáveis no Plano de Controle. Essencialmente, este plano abriga o núcleo de processamento do sistema proposto, incluindo seus três módulos principais: Assistentes Conversacionais, Resolutor de Conflitos e a Aplicação Inteligente (iApp) que atua primariamente como um construtor de sintaxe. O sistema proposto dispõe de dois assistentes conversacionais distintos e que admitem fluxos de processamento alternativos e independentes. Cada núcleo de Assistente Conversacional detém seu próprio par de *frontend* e *backend*, integrando ferramentas e *software* de código aberto dedicados a propósitos específicos de cada assistente.

O Assistente Conversacional baseado em Regras Pré-Definidas tem como alicerce Redes Neurais e prioriza a baixa latência no processamento das intenções através de um fluxo de processamento textual simplificado, focado em regras pré-definidas e operações mais eficientes. Para isso, utiliza um *frontend* integrado à API do aplicativo de mensagens instantâneas Telegram e um *backend* desenvolvido sobre o Rasa[1], um arcabouço para a criação de diálogos interativos entre usuários e máquinas. O Rasa opera com três componentes principais: o Agente, responsável pela intermediação do fluxo de solicitação-resposta; o Rasa NLU (*Natural Language Understanding*), responsável pela compreensão da linguagem natural; e o Rasa Core, responsável pelo gerenciamento do diálogo. Internamente o Rasa NLU desempenha a extração de entidades relevantes, classificação de intenções e recuperação de respostas através de rotinas personalizáveis de processamento de linguagem natural. O Rasa Core concentra-se na análise dos comandos processados de cada interação e na orientação das decisões de diálogo, prezando pela manutenção do contexto semântico da conversa. No Rasa NLU, o modelo DIET (*Dual Intent and Entity Transformer*) é uma rede neural treinada para classificar as intenções genéricas em intenções válidas, capacitando o sistema a lidar com entradas inéditas após treinamento. No Rasa Core, a abordagem de rede neural é desempenhada pelo modelo TED (*Transformer Embedding Dialogue*), capaz de decidir a próxima ação do assistente conversacional no diálogo com base nas interações passadas.

O Assistente Conversacional baseado em Modelos Generativos utiliza um modelo de linguagem em larga escala personalizado, ampliando consideravelmente sua abrangência interpretativa quando comparado ao assistente conversacional baseado em regras pré-definidas. Seu desenvolvimento é centrado na Open Web UI[2], uma plataforma de código aberto e extensível para interação com modelos de inteligência artificial, especialmente modelos de linguagem de grande escala (*Large Language Models* - LLMs). A plataforma disponibiliza um *frontend* próprio e intuitivo, capaz de receber intenções através de uma interface no formato de *chatbot*. Também oferece uma gama de funcionalidades para compor um *backend* personalizado. Dentre esses funcionalidades disponíveis, estão: (i) a escolha do modelo-base, i.e. modelos pré-treinados em grandes volumes de dados textuais permitindo que as aplicações que os utilizem sejam capacidade de lidar com textos inéditos e complexos; (ii) a descrição do *prompt*, ou seja, a

[1]Disponível em `https://rasa.com/`.
[2]Disponível em `https://openwebui.com/`.

instrução de entrada no modelo base, definindo textualmente a tarefa a ser desempenhada por ele. A qualidade do *prompt* influencia diretamente a qualidade da saída do modelo; (iii) o desenvolvimento de ferramentas (*tools*), que são *scripts* Python que expandem as capacidades de um LLM, permitindo a execução de rotinas ou ações específicas durante uma conversa, sem forçar seu uso. Em particular, o Assistente Conversacional baseado em Modelos Generativos adota como modelo-base o *Llama 3:8b*, um modelo de linguagem de código aberto desenvolvido pela Meta AI, composto por 8 bilhões de parâmetros. Além disso, o LLM do sistema proposto dispõe de duas ferramentas complementares. O Extrator de Parâmetros identifica os parâmetros essenciais para a interpretação da intenção, coletando-os e armazenando-os em uma estrutura JSON que contém informações como ação requisitada, entidades envolvidas, grandeza, unidade de medida, carimbo de tempo de início e fim da vigência da intenção. Já o Mapeador de Entidades, atua na interpretação de sinônimos e é essencialmente pautado na consulta a tabela de buscas, uma estrutura de dados contendo palavras relacionadas à palavra alvo dentro de um contexto semântico ou técnico. Tal tabela de buscas pode ser populada correlacionando diversos sinônimos possíveis dos termos relacionados ao cenário de redes, em especial Open RAN.

Internamente, ambos os assistentes conversacionais executam, em diferentes graus de interpretação e precisão, um *pipeline* de processamento textual que envolve três procedimentos básicos: tokenização, correção ortográfica e remoção de ruídos textuais. A tokenização é um procedimento de segmentação textual onde o texto original contíguo é dividido em fragmentos de acordo com um caractere delimitador. Como resultado, obtém-se um conjunto de *tokens*, tais como frases, palavras, letras, termos ou quaisquer estruturas textuais originalmente separadas pelo caractere escolhido. A correção ortográfica aplicada considera a distância Levenshtein, métrica que calcula o número mínimo de operações necessárias para transformar o *token* desconhecido em seu correspondente mais próximo em um dicionário de palavras corretas ortograficamente. Nesse procedimento de remoção de ruídos textuais, desconsideram-se estruturas gramaticais e ortográficas irrelevantes para o entendimento da intenção, tais como artigos, pronomes, pontuações e alguns caracteres especiais. A principal vantagem consiste na diminuição da pluralidade de palavras utilizadas na composição da intenção processada, reduzindo a complexidade de processos subsequentes.

Após o processamento linguístico em quaisquer dos assistentes conversacionais, as intenções são encaminhadas ao Resolvedor de Conflitos, módulo responsável por identificar e mitigar potenciais conflitos com as políticas de rede em execução. A identificação de conflitos, realizada por algoritmos de aprendizado profundo LSTM (*Long Short-Term Memory*) e GRU (*Gated Recurrent Unit*), avalia se as intenções propostas podem ser aceitas ou devem ser rejeitadas devido a incompatibilidades com políticas já implementadas. Um aspecto crucial desse processo é a correta identificação do domínio afetado pela intenção, ou seja, a porção específica da rede que será alvo da ação. O Resolvedor de Conflitos garante que a execução das intenções seja precedida por uma validação, notificando o operador sobre potenciais mudanças na gestão da rede decorrentes da execução de intenções conflitantes e permitindo a identificação e alteração de parâmetros que estejam fora dos padrões estabelecidos nos SLAs. Por fim, as intenções devidamente processadas e validadas são submetidas ao módulo Construtor de Sintaxe, responsável por convertê-las em comandos executáveis e em conformidade com a sintaxe exigida pelo controlador.

No **Plano de Controle**, o controle de rede é realizado por dois componentes: ONOS (*Open Network Operating System*) e O-RAN RIC (*RAN Intelligent Controller*). O ONOS é um controlador SDN que fornece uma interface centralizada para controle de infraestrutura de rede. O O-RAN RIC fornece controle centralizado e inteligência avançada para otimizar funções da RAN por meio de microaplicativos. A inserção de intenções no ONOS pode ocorrer via diferentes interfaces, tanto pela API Rest, a qual permite a inserção programática de intenções utilizando a saída JSON gerada pelo módulo iApp, quanto pela própria interface de linha de comando (*Command Line Interface*- CLI). A execução das intenções no ONOS pode ser desempenhada de duas maneiras, dependendo do grau de especificidade da intenção inserida pelo operador. Caso a intenção detenha um caráter passivo, ou seja, não induza nenhuma mudança no estado da rede e vise apenas o monitoramento das condições atuais e passadas da rede, esta será executada utilizando comandos tradicionais do ONOS. Contudo, caso a intenção possua um caráter ativo, ou seja, provoque alterações na configuração ou operação da rede instruindo a modificação de algum parâmetro, topologia, dispositivo ou métrica da rede, ela pode ser direcionada ao *ONOS Intent Framework*. Algumas intenções de caráter ativo mais simples, como a adição de novos dispositivos finais, não são encaminhadas ao *Intent Framework*. As intenções pré-existentes no *Intent Framework* abrangem encaminhamento e QoS.

O **Plano de Encaminhamento** é responsável pelo encaminhamento de pacotes e pela execução de políticas de rede definidas nos planos superiores. O plano de encaminhamento inclui os UEs, o núcleo da RAN e elementos de encaminhamento, como comutadores. Em ambiente emulado, o plano de encaminhamento implementa o *Open vSwitch*[3] (OVS), um comutador virtualizado que implementa funções de encaminhamento em *software*, facilitando a conectividade entre diferentes componentes da rede virtualizada. Em ambiente real, o plano de encaminhamento é composto por comutadores P4 que executam o SONiC (*Software for Open Networking in the Cloud*), um sistema operacional de rede de código aberto projetado para ser executado em *hardware* de comutadores de rede, permitindo a implementação de um comutador programável. UEs são emulados usando a o *software* srsUE da solução srsRAN[4], que implementa a gNodeB (gNB). A comunicação entre UEs e gNB é tratada via ZeroMQ[5], uma biblioteca de mensagens que oferece comunicação eficiente e de baixa latência entre processos distribuídos, que permite emular a comunicação entre UE e gNB. Essa comunicação também pode ser realizada com UEs físicas. O Open5GS[6] é usado como núcleo da rede, fornecendo funções essenciais de rede, como AMF (*Access and Mobility Management Function*), responsável por gerenciar o acesso e a mobilidade dos UEs, e UPF (*User Plane Function*), responsável pelo correto encaminhamento dos pacotes dos usuários.

5. Resultados Experimentais e Discussão

A avaliação prévia do sistema AGIR [Garcês et al., 2024, de Oliveira et al., 2024] mostra que a proposta executa corretamente a tradução de intenções de alto nível, evidenciando que a latência para instalação das intenções é baixa e tem crescimento linear com o número de intenções instaladas. Este artigo foca a avaliação da eficácia do controle de

[3]Disponível em `https://www.openvswitch.org/`.
[4]Disponível em `https://github.com/srsran/`.
[5]Disponível em `https://zeromq.org/`.
[6]Disponível em `https://open5gs.org/`.

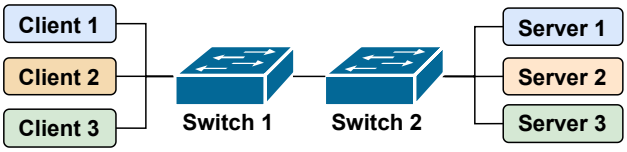


Figura 2. Topologia experimental emulada. Cada par cliente-servidor representa um fluxo de tráfego de uma categoria de serviço 5G.

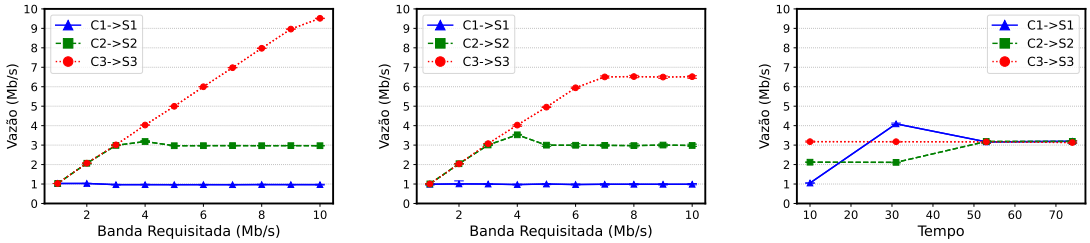
largura de banda considerando três fatias de rede que representam as categorias de serviço avançados 5G: eMBB, URLLC e mMTC. O eMBB é o serviço que mais demanda largura de banda, seguido pelo URLLC e pelo mMTC. A limitação de largura de banda resulta de uma intenção do operador que exige que a rede priorize uma classe de tráfego específica. A limitação da largura de banda evita que tráfego excessivo cause instabilidade na rede.

A avaliação é realizada como uma prova de conceito e utiliza duas abordagens distintas. Na primeira, o controle de largura de banda é implementado usando medidores OpenFlow. Os medidores são adicionados ao OVS e fluxos que apontam para esses medidores são criados. O medidor é configurado no OVS conectado aos clientes, pois o tráfego controlado flui do cliente para o servidor. Quando o tráfego excede o limite de largura de banda do medidor, os pacotes excedentes são descartados, garantindo que o tráfego não ultrapasse o limite estabelecido. Regras de fluxo são configuradas nos medidores, assegurando que cada dispositivo respeite os limites de largura de banda. O uso de medidores OpenFlow com regras de fluxo claras permite que os operadores de rede garantam a conformidade contínua com os SLAs, reduzindo o risco de penalidades contratuais ou degradação da experiência do usuário final.

Na segunda abordagem, utilizam-se filas implementadas no OVS. As filas permitem a priorização de diferentes fluxos, o que não é possível com os medidores OpenFlow do ONOS. As filas implementadas obedecem à disciplina HTB e cada fila corresponde a uma limitação específica de largura de banda. A HTB organiza o tráfego em uma estrutura hierárquica de filas, permitindo uma distribuição eficiente da largura de banda entre diferentes classes de tráfego, ao mesmo tempo que assegura controle preciso. Essas filas possuem largura de banda mínima garantida e máxima permitida e são associadas à porta de saída do OVS. Uma vez que os fluxos são criados e vinculados às filas, a HTB prioriza o tráfego com base na largura de banda mínima garantida.

A topologia experimental é apresentada na Figura 2. Três clientes estão conectados a três servidores por meio de dois OVSs. Cada comunicação cliente-servidor representa uma categoria de serviço avançado 5G: mMTC do Cliente 1 ($C1$) para o Servidor 1 ($S1$), URLLC do Cliente 2 ($C2$) para o Servidor 2 ($S2$) e eMBB do Cliente 3 ($C3$) para o Servidor 3 ($S3$). Para simular um ambiente com recursos limitados, o enlace entre os dois comutadores virtuais é restrito a 10 Mb/s. Utiliza-se o iperf[7] como gerador de tráfego. Assim, cada cliente gera um tráfego TCP sintético de taxa de bits constante. A vazão máxima alcançada por cada fluxo é medida. Todos os experimentos são realizados em uma máquina virtual executando Ubuntu 22.04 LTS, equipada com um Processador Virtual QEMU versão 2.5+ 3.2 GHz, 4 GB de RAM e 40 GB de armazenamento. Os resultados são apresentados com intervalos de confiança de 95%.

[7]Disponível em `https://iperf.fr/`.

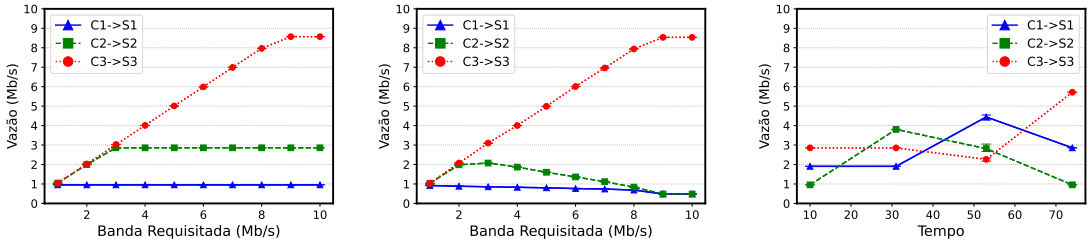


(a) Fluxos individuais, alocação de banda estática.

(b) Fluxos simultâneos, alocação de banda estática.

(c) Fluxos simultâneos, alocação de banda dinâmica.

Figura 3. Taxa de transferência máxima alcançada em função da largura de banda solicitada. A largura de banda permitida para cada fluxo é atribuída aos Medidores OpenFlow. Fluxos individuais recebem a largura de banda solicitada até o limite permitido. Fluxos simultâneos compartilham a capacidade do enlace e o fluxo com maior demanda não consegue atender ao seu SLA.



(a) Fluxos individuais, alocação de banda estática, diferentes prioridades

(b) Fluxos simultâneos, alocação de banda estática, diferentes prioridades

(c) Fluxos simultâneos, alocação de banda dinâmica, diferentes prioridades

Figura 4. Taxa de transferência máxima alcançada em função da largura de banda solicitada. A largura de banda permitida para cada fluxo é atribuída às filas HTB. Fluxos individuais recebem a largura de banda solicitada até o limite permitido. Fluxos simultâneos compartilham a capacidade do enlace e o fluxo com maior demanda é priorizado, permitindo que atenda ao seu SLA.

Considerando ambas as abordagens, avalia-se a vazão alcançada por cada fluxo quando estão presentes individualmente na rede e quando coexistem simultaneamente. Também avalia-se um cenário dinâmico, no qual as regras de largura de banda dos fluxos mudam em tempo de execução. A avaliação dos fluxos individuais visa verificar se o limite de largura de banda é respeitado. Já a avaliação dos fluxos simultâneos verifica o comportamento dos fluxos com e sem priorização. Ao avaliar o cenário com regras dinâmicas de largura de banda é possível analisar a variação da vazão de cada fluxo ao longo do tempo, com e sem priorização.

Usando os medidores OpenFlow, a limitação de cada fluxo é a seguinte: $C1 \rightarrow S1$ é limitado a 1 Mb/s, $C2 \rightarrow S2$ a 3 Mb/s e $C3 \rightarrow S3$ a 9,5 Mb/s. Verifica-se a limitação imposta aumentando a largura de banda solicitada de cada fluxo. A Figura 3 mostra os resultados. Quando cada fluxo existe isoladamente na rede, à medida que a largura de banda solicitada aumenta, a vazão de cada fluxo tende ao limite de largura de banda estabelecido, como mostrado na Figura 3(a). Como esperado, a capacidade total do enlace afeta a vazão alcançada pelos fluxos quando eles coexistem simultaneamente. A Figura 3(b) mostra que o Fluxo 1, de $C1$ para $S1$, não consegue atingir a largura de banda

solicitada, mesmo quando está abaixo do limite de largura de banda, devido à capacidade do enlace. Assim, o fluxo de maior prioridade, de $C1$ para $S1$, não atende ao SLA, enquanto os fluxos de menor prioridade, de $C2$ para $S2$ e $C3$ para $S3$, alcançam suas larguras de banda solicitadas. A Figura 3(c) mostra o resultado do cenário dinâmico com fluxos simultâneos. Inicialmente, as limitações são de 1 Mb/s, 2 Mb/s e 3 Mb/s para os Fluxos 1, 2 e 3, de $C1 \rightarrow S1$, $C2 \rightarrow S2$ e $C3 \rightarrow S3$, respectivamente. Em 10 segundos, todos os fluxos atingem a vazão máxima limitada pelos seus respectivos limites de largura de banda. Após alguns segundos, o limite de largura de banda do Fluxo 1 é aumentado para 4 Mb/s, resultando em uma vazão de 4 Mb/s em 30 segundos. Em seguida, o Fluxo 2 é limitado a 5 Mb/s, o que faz com que a largura de banda total exceda a capacidade do enlace. Como não há algoritmo priorizando o tráfego, a largura de banda é dividida igualmente entre os fluxos.

A abordagem que utiliza filas HTB permite priorizar os fluxos. A Figura 4 mostra a vazão máxima alcançada considerando a largura de banda solicitada, a alocação de largura de banda e as prioridades dos fluxos. A limitação de largura de banda de cada fluxo é a seguinte: $C1 \rightarrow S1$ é limitado a 1 Mb/s, $C2 \rightarrow S3$ a 3 Mb/s e $C3 \rightarrow S3$ a 9,5 Mb/s. A prioridade é definida de acordo com o limite de largura de banda e, assim, $C3 \rightarrow S3$ tem a maior prioridade, representando eMBB, $C2 \rightarrow S2$ tem prioridade média, representando URLLC, e $C1 \rightarrow S1$ tem a menor prioridade, representando mMTC. Quando os fluxos individuais existem isoladamente na rede, a vazão máxima é restrita pelo limite de largura de banda de cada fluxo, como mostra a Figura 4(a). No caso de fluxos simultâneos, à medida que as larguras de banda solicitadas aumentam, as vazões também aumentam até o limite de largura de banda se a soma das vazões for menor que a capacidade do enlace. Caso contrário, a capacidade do enlace também limita a vazão máxima. Devido à priorização, os fluxos de menor prioridade têm suas vazões reduzidas para que o fluxo de maior prioridade possa alcançar seu SLA. Isso é mostrado na Figura 4(b). Assim, o fluxo de maior prioridade, $C3 \rightarrow S3$, exibe uma forte correlação entre a largura de banda solicitada e a vazão medida.

Observa-se que o mecanismo de QoS favorece o Fluxo 3, limitando os recursos consumidos pelos outros fluxos. Em contraste, o comportamento entre os diferentes pares cliente-servidor varia ao não aplicar a priorização e limitação baseada em filas. No cenário dinâmico, cada fluxo é inicialmente limitado da seguinte forma: $C1 \rightarrow S1$ é limitado a 1 Mb/s, $C2 \rightarrow S2$ a 2 Mb/s e $C3 \rightarrow S3$ a 3 Mb/s. Em seguida, o fluxo $C2 \rightarrow S2$ torna-se mais prioritário ao aumentar seu limite de largura de banda para 4 Mb/s. Posteriormente, $C1 \rightarrow S1$ torna-se o fluxo de maior prioridade ao configurar seu limite de largura de banda para 5 Mb/s. Finalmente, $C3 \rightarrow S3$ retorna como o fluxo de maior prioridade ao aumentar seu limite de largura de banda para 6 Mb/s. A Figura 4(c) mostra a vazão de cada fluxo ao longo do tempo. Quando a soma das larguras de banda solicitadas ultrapassa a capacidade do enlace, o algoritmo HTB distribui a largura de banda de acordo com as prioridades dos fluxos. Isso ocorre quando os limites de largura de banda dos fluxos $C2 \rightarrow S2$ e $C3 \rightarrow S3$ aumentam para 5 Mb/s e 6 Mb/s, respectivamente. Assim, os Fluxos 2 e 3, em aproximadamente 52 e 75 segundos, alcançam a vazão máxima em troca da redução das vazões dos outros fluxos. Esse comportamento difere da abordagem com medidores OpenFlow, que favorece a equidade entre os fluxos.

6. Conclusão

As redes móveis de próxima geração, B5G/6G, enfrentam desafios significativos para garantir a QoS em um ambiente complexo e distribuído, especialmente no contexto de arquiteturas Open RAN. A crescente demanda por serviços avançados, como eMBB, URLLC e mMTC, impõe a necessidade de alocação eficiente de recursos que assegure o cumprimento dos Acordos de Nível de Serviço (SLAs) e a coexistência de diferentes fatias de rede. Este artigo propôs um sistema de gerenciamento de fatiamento de rede baseado em intenções, no qual os operadores podem controlar a alocação de recursos por meio de entradas expressas em linguagem natural, automaticamente convertidas em instruções aplicadas no plano de encaminhamento da rede. Os resultados mostram que o sistema proposto garante, de forma eficiente, o cumprimento dos SLAs em diferentes serviços B5G/6G, utilizando tanto medidores OpenFlow quanto filas HTB. Os experimentos emulados validaram a capacidade do sistema de ajustar dinamicamente os recursos da rede, garantindo que serviços de maior prioridade, como eMBB, recebam os recursos necessários ao utilizar filas HTB.Como trabalhos futuros, pretendemos estender o gerenciamento baseado em intenções para abranger tanto os recursos da rede de núcleo e quanto os da RAN. Além disso, planeja-se avaliar o desempenho do modelo de linguagem em larga escala adotado como um dos núcleos do assistente conversacional para aumentar a precisão e a flexibilidade do sistema.

Referências

Abbas, K., Afaq, M., Ahmed Khan, T., Rafiq, A. e Song, W.-C. (2020). Slicing the core network and radio access network domains through intent-based networking for 5g networks. *Electronics*, 9(10):1710.

Abbas, K., Khan, T. A., Afaq, M. e Song, W.-C. (2021). Network slice lifecycle management for5G mobile networks: An intent-based networking approach. *IEEE Access*, 9:80128–80146.

Afolabi, I., Taleb, T., Samdanis, K., Ksentini, A. e Flinck, H. (2018). Network slicing and softwarization: A survey on principles, enabling technologies, and solutions. *IEEE Communications Surveys & Tutorials*, 20(3):2429–2453.

Aklamanu, F., Randriamasy, S., Renault, E., Latif, I. e Hebbar, A. (2018). Intent-based real-time 5g cloud service provisioning. Em *2018 IEEE Globecom Workshops (GC Wkshps)*, p. 1–6. IEEE.

Da Costa, A. M. e Murillo, L. M. C. (2023). Integration of network slice controller for enhanced intent-based networking in 5G/6G networks. Em *Proceedings of the 18th Workshop on Mobility in the Evolving Internet Architecture*, MobiArch '23, p. 31–36, New York, NY, USA. Association for Computing Machinery.

de Oliveira, N. R., Medeiros, D. S., Moraes, I. M., Andreonni, M. e Mattos, D. M. (2024). Towards intent-based management for open radio access networks: an agile framework for detecting service-level agreement conflicts. *Annals of Telecommunications*, p. 1–14.

de Oliveira, N. R., Moraes, I. M., de Medeiros, D. S. V., Lopez, M. A. e Mattos, D. M. (2023). An agile conflict-solving framework for intent-based management of service

level agreement. Em *2023 2nd International Conference on 6G Networking (6GNet)*, p. 1–8. IEEE.

Elayoubi, S. E., Jemaa, S. B., Altman, Z. e Galindo-Serrano, A. (2019). 5g ran slicing for verticals: Enablers and challenges. *IEEE Communications Magazine*, 57(1):28–34.

Faruque, M. A. (2021). A review study on "5g nr slicing enhancing iot & smart grid communication". Em *2021 12th International Renewable Engineering Conference (IREC)*, p. 1–4.

Filali, A., Mlika, Z., Cherkaoui, S. e Kobbane, A. (2022). Dynamic sdn-based radio access network slicing with deep reinforcement learning for urllc and embb services. *IEEE Transactions on Network Science and Engineering*, 9(4):2174–2187.

Garcês, J., Oliveira, N., Watanabe, J., Tanaka, R., Arruda, D., Leite, B., Galdino, C., Couto, R., Moraes, I., Medeiros, D. e Mattos, D. (2024). Gerenciamento baseado em intenção para a Open RAN: Automação inteligente de configuração de redes via chatbot. Em *Anais do XXIX Workshop de Gerência e Operação de Redes e Serviços*, p. 210–223, Porto Alegre, RS, Brasil. SBC.

Giorgetti, A., Scano, D. e Valcarenghi, L. (2021). Guaranteeing slice performance isolation with sdn. *IEEE Communications Letters*, 25(11):3537–3541.

Habib, M. A., Zhou, H., Iturria-Rivera, P. E., Elsayed, M., Bavand, M., Gaigalas, R., Ozcan, Y. e Erol-Kantarci, M. (2023). Intent-driven intelligent control and orchestration in o-ran via hierarchical reinforcement learning. Em *2023 IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, p. 55–61. IEEE.

Infiesta, J. T., Guimarães, C., Contreras, L. M. e de la Oliva, A. (2020). Ganso: Automate network slicing at the transport network interconnecting the edge. Em *2020 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, p. 161–166. IEEE.

Rezende, P., Curado, M. e Madeira, E. (2023). NS-ENFORCER: Enforcing Network Slicing on Radio Access Networks. *Journal of Network and Systems Management*, 31(2):30.

Rost, P., Mannweiler, C., Michalopoulos, D. S., Sartori, C., Sciancalepore, V., Sastry, N., Holland, O., Tayade, S., Han, B., Bega, D., Aziz, D. e Bakker, H. (2017). Network slicing to enable scalability and flexibility in 5g mobile networks. *IEEE Communications Magazine*, 55(5):72–79.

Yeh, S.-p., Bhattacharya, S., Sharma, R. e Moustafa, H. (2023). Deep learning for intelligent and automated network slicing in 5g open ran (oran) deployment. *IEEE Open Journal of the Communications Society*.

Zhang, J., Wei, H., Gao, D., Xia, N., Wang, D., Yan, S. e Liu, X. (2023). Intent-driven ran slice orchestration: A multi-agent deep reinforcement learning based approach. Em *2023 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, p. 1–6.